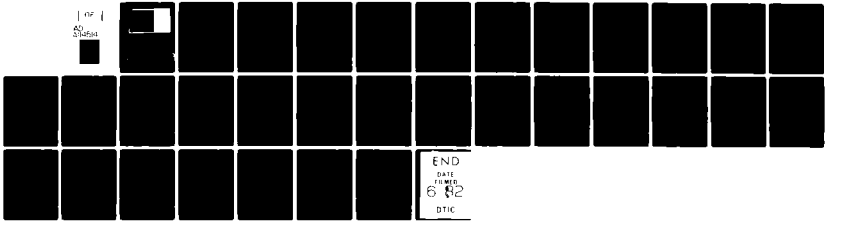
AD-A114 514 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
THE CENTRAL ROLE OF THE PROPENSITY SCORE IN OBSERVATIONAL STUDI--ETC(U)
DEC 81 P R ROSENBAUM, D B RUBIN DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2305 NL



END
DATE FILMED
6 82
DTIC

MRC Technical Summary Report #2305

THE CENTRAL ROLE OF THE PROPENSITY SCORE IN OBSERVATIONAL STUDIES FOR CAUSAL EFFECTS

Paul R. Rosenbaum and Donald B. Rubin

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

December 1981

(Received September 14, 1981)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Cancer Institute
9000 Rockville Pike
Bethesda, MD 20205

Health Resources Administration
3700 East-West Highway
Hyattsville, MD 20782

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

THE CENTRAL ROLE OF THE PROPENSITY SCORE IN OBSERVATIONAL STUDIES FOR CAUSAL EFFECTS

Paul R. Rosenbaum and Donald B. Rubin

Technical Summary Report #2305

December 1981

ABSTRACT

The results of observational studies are often disputed because of nonrandom treatment assignment. For example, patients at greater risk may be overrepresented in some treatment groups. This paper discusses the central role of "propensity scores" and "balancing scores" in the analysis of observational studies. The propensity score is the (estimated) conditional probability of assignment to a particular treatment given a vector of observed covariates. Both large and small sample theory show that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Applications include: (1) matched sampling on the univariate propensity score which is equal percent bias reducing under more general conditions than required for discriminant matching, (2) multivariate adjustment by subclassification on balancing scores where the same subclasses are used to estimate treatment effects for all outcome variables and in all subpopulations, and (3) visual representation of multivariate adjustment by a two-dimensional plot.

AMS (MOS) Subject Classifications: 62P99, 62J05, 62F12, 62F11

Key Words: Observational studies; covariance adjustment; subclassification; matching.

Work Unit Number 4 - Statistics and Probability

This work was partially supported by Health Resources Administration contract HRA 230-76-0300, partially by the Division of Statistics and Applied Mathematics, Office of Radiation Programs, U.S. Environmental Protection Agency, partially by grant P30-CA-14520 from the National Cancer Institute to the Wisconsin Clinical Cancer Center, and partially sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

ACKNOWLEDGEMENT

The authors wish to acknowledge valuable discussions with Arthur P. Dempster and Roderick J. A. Little on the subject of this paper.

THE CENTRAL ROLE OF THE PROPENSITY SCORE IN OBSERVATIONAL STUDIES FOR CAUSAL EFFECTS

Paul R. Rosenbaum and Donald B. Rubin

1. DEFINITIONS

An experiment is defined as a comparison of several treatments, any one of which may be given to or withheld from any of N units (e.g., medical patients) under study. Inferences about the effects of treatments involve speculations about the effect one treatment would have had on a unit which, in fact, received some other treatment. In a series of papers, Rubin (e.g., 1978) formalized this concept in a way consistent with that traditionally used in the literature of experimental design (e.g., Fisher (1953) and Kempthorne (1952)). Suppose there are only two treatments 1 and 2. In principle, the i^{th} of the N units under study has both a response r_{1i} that would have resulted if it had received treatment 1, and a response r_{0i} that would have resulted if it had received treatment 2. In this formulation, causal effects are comparisons of r_{1i} and r_{0i} (e.g. $r_{1i} - r_{0i}$ or r_{1i}/r_{0i}). Since each unit receives only one treatment, either r_{1i} or r_{0i} is observed, but not both, so comparisons of r_{1i} and r_{0i} imply some degree of speculation. In a sence, estimating the causal effects of treatments is a missing data problem, since either r_{1i} or r_{0i} is missing.

The above formulation contains some implicit assumptions. For example, the response r_{ti} of unit i to treatment t might depend on the treatment given to unit j, if for example, they compete for resources. Such a

This work was partially supported by Health Resources Administration contract HRA 230-76-0300, partially by the Division of Statistics and Applied Mathematics, Office of Radiation Programs, U.S. Environmental Protection Agency, partially by grant P30-CA-14520 from the National Cancer Institute to the Wisconsin Clinical Cancer Center, and partially sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

situation complicates not only the analysis, but the definition of the causal effect as well. Such problems are not considered in this paper. For a fuller discussion, see Cox (1958, chapter 2) or Rubin (1978, section 2.3).

In this paper, the N units in the study are viewed as a simple random sample from some population, and the average treatment effect is defined as

$$E(r_1) - E(r_0) \tag{1.1}$$

where $E(\cdot)$ denotes expectation in the population. In large randomized experiments, the results in the two treatment groups may often be directly compared because the units in the two treatment groups are likely to be similar, whereas in nonrandomized experiments, such direct comparisons may be misleading because units exposed to one treatment generally differ systematically from the units exposed to the other treatment. Cox (1981, p. 291) has observed that there is a need for further discussion of observational studies with particular emphasis on bias isolation and removal.

For the i^{th} patient of N patients in the study $(i=1,\ldots,N)$, let z_i be the indicator for treatment assignment, with $z_i = 1$ if unit i is assigned to the experimental treatment, and $z_i = 0$ if unit i is assigned to the control treatment. Let $\underline{x}_i$ be a vector of observed pretreatment measurements or covariates for the i^{th} unit; all of the measurements in $\underline{x}$ were made prior to treatment assignment, but $\underline{x}$ may not include all covariates used to make treatment assignments.

Suppose each unit can be assigned a scalar "balancing" score $b(\underline{x})$ such that, at each value of the balancing score, the distribution of the observed covariates $\underline{x}$ is the same for the treated and control units; that is, suppose $b(\underline{x})$ exists such that, in Dawid's (1978) notation,

$$z \perp\!\!\!\perp \underline{x} \mid b(\underline{x}) \quad .$$

Then, at each value of the balancing score, the difference between treatment and control means on the response r is unconfounded with $\underline{x}$, although it may be confounded with unobserved covariates.

We prove that such a balancing score always exists, and then show that easily obtained estimates of the balancing score behave like balancing scores; indeed, in sections 3.3 and 4.2 we find that an estimated balancing score can produce greater sample balance than population balancing score. Moreover, in section 4 we see that common methods of adjustment in observational studies -- including covariance adjustment, and discriminant matching (Cochran and Rubin, 1973) -- implicitly adjust for an estimated balancing score.

In order to motivate formally adjustment for a balancing score, we must consider the sampling distribution of treatment assignments. Let the conditional probability of assignment to treatment one, given the covariates, be denoted by

$$e(\underline{x}) = p(z = 1 \mid \underline{x}) \tag{1.2}$$

where we assume $p(z_1,\ldots,z_n \mid \underline{x}_1,\ldots,\underline{x}_n) = \prod_{i=1}^{n} e(\underline{x}_i)^{z_i}(1 - e(\underline{x}_i))^{1-z_i}$. Although this strict independence assumption is not essential, it simplifies notation and discussion. The function $e(\underline{x})$ is called the propensity score, that is, the propensity towards exposure to treatment one given the observed covariates $\underline{x}$.

Randomized and nonrandomized trials differ in two distinct ways. First, in a randomized trial, z_i has a distribution determined by a known random mechanism; therefore, in particular, the propensity score is a known function: there exists one accepted specification for $e(\underline{x})$. In a nonrandomized experiment, the propensity score function is almost always unknown: there is not one accepted specification for $e(\underline{x})$; however, $e(x)$ may be estimated from

observed data, perhaps using a model such as a logit model. To a Bayesian, estimates of these probabilities are posterior predictive probabilities of assignment to treatment 1 for a unit with vector $\underline{x}$ of covariates.

The second way randomized trials differ from nonrandomized trials is that, in a randomized trial, $\underline{x}$ is known to contain all covariates that are both used to assign treatments and possibly related to the response (r_{1i}, r_{0i}). More formally, in a randomized trial, treatment assignment z_i and response (r_{1i}, r_{0i}), are known to be conditionally independent given $\underline{x}_i$,

$$(r_{1i}, r_{0i}) \perp\!\!\!\perp z_i \mid \underline{x}_i \quad . \tag{1.3}$$

Condition (1.3) is usually not known to hold in a nonrandomized experiment. Generally, we shall say treatment assignment is <u>strongly ignorable given a vector of covariates</u> $\underline{v}$ if

$$(r_{1i}, r_{0i}) \perp\!\!\!\perp z_i \mid \underline{v}_i \quad .$$

For brevity, when treatment assignment is strongly ignorable given the observed covariates $\underline{x}$ (i.e., when (1.3) holds), we shall say simply that <u>treatment assignment is strongly ignorable</u>. (Note that if treatment assignment is strongly ignorable, then it is ignorable in Rubin's (1978) sense, which only requires that the probabilities be evaluated at observed outcomes; however, the converse is not true since strongly ignorable implies the relationship among probabilities must hold for all possible values of the random variables.)

2. LARGE SAMPLE THEORY

This section presents four theorems whose conclusions may be summarized as follows.

(1) The propensity score is a balancing score.

(2) Any score which is "finer" than a propensity score is a balancing score.

(3) If treatment assignment is strongly ignorable given $\underline{x}$, then it is strongly ignorable given any balancing score.

(4) At any value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score <u>if</u> treatment assignment is strongly ignorable.

The results of this section treat $e(\underline{x})$ as known, and are therefore applicable to large samples. The effects of estimating $e(\underline{x})$ in small samples are considered in section 3.

<u>Theorem 1</u>

Treatment assignment and the observed covariates are conditionally independent given the propensity score, that is

$$z \perp\!\!\!\perp \underline{x} \mid e(\underline{x}) \quad .$$

The above theorem is a special case of the Theorem 2, and so no separate proof is given. However, Cochran and Rubin (1973) proved this general result in the special case of multivariate normal covariates $\underline{x}$; the result holds regardless of the distribution of $\underline{x}$.

<u>Theorem 2</u>

Let $\underline{b}^*(\underline{x})$ be a (possibly vector valued) function of $\underline{x}$ which is finer than $e(\underline{x})$ in the same sense that $e(\underline{x}) = f(\underline{b}^*(\underline{x}))$ for some function $f(\cdot)$. Then

$$z \perp\!\!\!\perp \underline{x} \mid \underline{b}^*(\underline{x}) \quad . \tag{2.1}$$

In particular, if $\underline{b}^*(\underline{x})$ is scalar valued, then (2.1) asserts that $\underline{b}^*(\underline{x})$ is a balancing score.

Proof: It is sufficient to show

$$p(z=1 \mid \underline{x}) = p(z=1 \mid \underline{b}^*(\underline{x})) \quad .$$

Recall that, by definition, $e(\underline{x}) = p(z=1 \mid \underline{x})$. Now

$$\begin{aligned} & p(z=1 \mid \underline{b}^*(\underline{x}) = \underline{c}) \\ &= \int_{\underline{x}:\underline{b}^*(\underline{x})=\underline{c}} p(z=1 \mid \underline{x})\, p(\underline{x} \mid \underline{b}^*(\underline{x}) = \underline{c})d\underline{x} \\ &= e(\underline{x}) \int_{\underline{x}:\underline{b}^*(\underline{x})=\underline{c}} p(\underline{x} \mid \underline{b}^*(\underline{x}) = \underline{c})d\underline{x} \\ &= e(\underline{x}) \\ &= p(z=1 \mid \underline{x}) \end{aligned}$$

as required. //

Theorem 1 implies that if a subclass of units or a matched treatment-control pair is homogeneous in $e(\underline{x})$, then the treated and control units in that subclass or matched pair will have the same distribution of $\underline{x}$. Theorem 2 implies that if subclasses or matched treatment-control pairs are homogeneous in both $e(\underline{x})$ and certain chosen components of $\underline{x}$, it is still reasonable to expect balance on the other components of $\underline{x}$ within these refined subclasses or matched pairs. The practical importance of Theorem 2 beyond Theorem 1 arises because it is sometimes advantageous to subclassify or match not only for $e(\underline{x})$, but for other components of $\underline{x}$ as well; in particular, such a refined procedure may be used to obtain estimates of the average treatment effect in subpopulations defined by components of $\underline{x}$, (e.g., males, females).

Theorem 3, below, is the key result for showing that if treatment assignment is strongly ignorable then adjustment for a balancing score $b(\underline{x})$ is sufficient to produce unbiased estimates of the average treatment effect (1.1).

Theorem 3

If treatment assignment is strongly ignorable given $\underline{x}$, then it is strongly ignorable given the balancing score $b(\underline{x})$; that is,

$$(r_1, r_0) \perp\!\!\!\perp z \mid \underline{x}$$

implies

$$(r_1, r_0) \perp\!\!\!\perp z \mid b(\underline{x}) \quad .$$

Proof: By assumption

$$p(r_1, r_0, z, \underline{x}) = p(r_1, r_0 \mid \underline{x})p(z \mid \underline{x})p(\underline{x})$$

which equals

$$p(r_1, r_0 \mid \underline{x})p(z \mid b(\underline{x}))p(\underline{x}) \quad ,$$

since $b(\underline{x})$ is a balancing score. Then

$$\begin{aligned} & p(r_1, r_0, z \mid b(\underline{x}) = c) \\ &= \int_{\underline{x}:b(\underline{x})=c} p(r_1, r_0 \mid \underline{x})p(z \mid b(\underline{x}) = c)p(\underline{x})d\underline{x} \\ &= p(z \mid b(\underline{x}) = c) \int_{\underline{x}:b(\underline{x})=c} p(r_1, r_0 \mid \underline{x})p(\underline{x})d\underline{x} \\ &= p(z \mid b(\underline{x}) = c)p(r_1, r_0 \mid b(\underline{x}) = c) \end{aligned}$$

as required. //

We are now ready to relate balancing scores and ignorable treatment assignment to the estimation of treatment effects.

The response r_t to treatment t is observed only if the unit receives treatment t (i.e., $z = t$). Thus, if a randomly selected treated unit ($z = 1$) is compared to a randomly selected control unit ($z = 0$), the expected difference in response is

$$E(r_1 \mid z = 1) - E(r_0 \mid z = 0) \quad . \tag{2.2}$$

Expression (2.2) does not equal (1.1) in general because the available samples are not from the marginal distribution of r_t, but rather from the conditional distribution of r_t given $z = t$. In other words, in general, randomly selected units cannot act as controls for one another; i.e. the expected difference in their responses does not generally equal the average treatment effect.

Suppose a specific value of the vector of covariates $\underline{x}$ is randomly sampled from the entire population of units--both treated and control units together--and then a treated unit and a control unit are found both having this value for the vector of covariates. In this two step sampling process, the expected difference in response is

$$E_{\underline{x}}[E(r_1 \mid \underline{x}, z = 1) - E(r_0 \mid \underline{x}, z = 0)] \quad , \tag{2.3}$$

where $E_{\underline{x}}$ denotes expectation with respect to the distribution of $\underline{x}$ in the entire population of units. If treatment assignment is strongly ignorable, that is if (1.3) holds, then (2.3) equals

$$E_{\underline{x}}[E(r_1 \mid \underline{x}) - E(r_0 \mid \underline{x})] \quad ,$$

which does equal the average treatment effect (1.1). In other words, with strongly ignorable treatment assignment, two units with the same $\underline{x}$ but different treatments can act as controls for one another; i.e., the expected difference in their responses equals the average treatment effect. This formal observation is due to Rubin (1977), although it is implicit in earlier discussions of experimental design (e.g., Cox, 1958, Chapter 2).

Now suppose a value of a balancing score $b(\underline{x})$ is sampled from the entire population of units and then a treated unit and a control unit are sampled from all patients having this value of $b(\underline{x})$, but perhaps different values of $\underline{x}$. Given strongly ignorable treatment assignment, it follows from Theorem 3 that

$$E(r_1 \mid b(\underline{x}), z = 1) - E(r_0 \mid b(\underline{x}), z = 0)$$

$$= E(r_1 \mid b(\underline{x})) - E(r_0 \mid b(\underline{x}))$$

from which it follows that

$$\begin{aligned} & E_{b(\underline{x})}[E(r_1 \mid b(\underline{x}), z = 1) - E(r_0 \mid b(\underline{x}), z = 0)] \\ & \quad = E_{b(\underline{x})}[E(r_1 \mid b(\underline{x})) - E(r_0 \mid b(\underline{x}))] \\ & \quad = E(r_1 - r_0) \end{aligned} \tag{2.4}$$

where $E_{b(\underline{x})}$ denotes expectation with respect to the distribution of $b(\underline{x})$ in the entire population. In words, under strongly ignorable treatment assignment, units with the same value of the balancing score $b(\underline{x})$ but different treatments can act as controls for each other, in the sense that the expected difference in their responses equals the average treatment effect.

The above argument has established the following theorem.

Theorem 4.

Suppose treatment assignment is strongly ignorable. Suppose further that a group of patients is sampled using $\underline{x}$ such that (1) $b(\underline{x})$ is constant for all patients in the group, and (2) at least one patient received each treatment. Then, for these patients, the expected difference in treatment means equals the average treatment effect at that value of $b(\underline{x})$; that is,

$$E(r_1 \mid b(\underline{x}), z = 1) - E(r_0 \mid b(\underline{x}), z = 0)$$

$$= E(r_1 - r_0 \mid b(\underline{x})) \quad .$$

3. PRACTICAL ISSUES IN THE USE OF BALANCING SCORES

In practice, homogeneous subclasses and exact matches on balancing scores are difficult to obtain, and moreover the balancing scores must be estimated from the data. This section considers three practical issues: the effects of imperfect control for balancing scores; models for the propensity score $e(\underline{x})$; the effects in small samples of using an estimated balancing score.

3.1 Consequences of Imperfect Control for Balancing Scores

Since it is generally difficult in practice to find treated and control units for comparison with exactly the same value of the propensity score, units with similar but not identical values of the propensity score may be required. Theorem 5 below shows that if $\tilde{e}(\underline{x})$ is a close approximation to $e(\underline{x})$ then subclassification on $\tilde{e}(\underline{x})$ almost balances $\underline{x}$. For example, $e(\underline{x})$ might be a continuous function of continuous $\underline{x}$, whereas $\tilde{e}(\underline{x})$ might be a discrete approximation to $e(\underline{x})$ used to define a few subclasses within which $\tilde{e}(\underline{x})$ is constant.

Theorem 5

Suppose $|\tilde{e}(\underline{x}) - e(\underline{x})| < \varepsilon$ for all $\underline{x}$. Then

$$|p(z, \underline{x}|\tilde{e}(\underline{x})) - p(z|\tilde{e}(\underline{x}))p(\underline{x}|\tilde{e}(\underline{x}))| < 2\varepsilon \text{ for all } \underline{x} .$$

Proof: Since

$$\begin{aligned} &|p(A|B,C) - p(A|C)| \\ &\quad \geq |p(A|B,C) - p(A|C)|\, p(B|C) \\ &\quad = |p(A,B|C) - p(A|C)\, p(B|C)| \end{aligned}$$

it is sufficient to show that

$$|p(z=1|\underline{x}) - p(z=1|\tilde{e}(\underline{x}))| < 2\varepsilon .$$

Now, pick an $\underline{x}$, and let c be the value of $\tilde{e}(\underline{x})$. Then

$$\begin{aligned} &|p(z=1|\underline{x}) - p(z=1|\tilde{e}(\underline{x}) = c)| \\ &\quad = |e(\underline{x}) - \int_{\underline{v}:\tilde{e}(\underline{v})=c} p(z=1|\underline{v})p(\underline{v}|\tilde{e}(\underline{v}) = c)d\underline{v}| \end{aligned}$$

$$= \left| e(\underline{x}) - \tilde{e}(\underline{x}) + \int_{\underline{v}:\tilde{e}(\underline{v})=c} [e(\underline{v}) - \tilde{e}(\underline{v})] p(\underline{v} | \tilde{e}(\underline{v}) = c) d\underline{v} \right|$$

$< 2\varepsilon$ as required. //

3.2 The Form of the Propensity Score Under Various Models

Often the propensity scores $e(\underline{x})$ must be estimated from available data. Therefore, it is convenient to note that the propensity score has familiar forms under certain familiar models; in particular, the propensity score can often be modelled using an appropriate logit model (Cox (1970)) or discriminant score.

Clearly,

$$e(\underline{x}) = p(z = 1 \mid \underline{x}) = \frac{p(z = 1)p(\underline{x} \mid z = 1)}{p(z = 1)p(\underline{x} \mid z = 1) + p(z = 2)p(\underline{x} \mid z = 0)} .$$

Elementary manipulations establish the following facts.

1. If $p(\underline{x} \mid z = t) = N_p(\underline{\mu}_t, \Sigma)$ then $e(\underline{x})$ is a monotone function of the linear discriminant $\underline{x}^T \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$. Therefore, stratification or matching on $e(\underline{x})$ includes discriminant matching as a special case. This method was first proposed by Cochran and Rubin (1973), and was studied further in Rubin (1976, 1979, 1980).

2. If $p(\underline{x} \mid z = t)$ is a polynomial exponential family distribution, i.e., if

$$p(\underline{x} \mid z = t) = h(\underline{x}) \exp(P_t(\underline{x}))$$

where $P_t(\underline{x})$ is a polynomial in $\underline{x}$ of degree k, say, then $e(\underline{x})$ obeys a polynomial logit model

$$\log_e \frac{e(\underline{x})}{1 - e(\underline{x})} = \log_e \frac{p(z = 1)}{1 - p(z = 1)} + P_1(\underline{x}) - P_2(\underline{x})$$

$$= \log_e \frac{p(z = 1)}{1 - p(z = 1)} + Q(\underline{x})$$

where $Q(\underline{x})$ is a degree k polynomial in $\underline{x}$.

This polynomial exponential family includes the linear exponential family (resulting in a linear logit model for $e(\underline{x})$), the quadratic exponential family described in Dempster (1971), and the binary data model described in Cox (1975).

3.3 Small Sample Theory for Discrete $\underline{x}$

This section demonstrates that subclassification on a sample estimate $\hat{e}(\underline{x})$ of the propensity score $e(\underline{x})$ produces sample balance, that is balance in terms of the sample or empirical distributions. Although the theorems of this section are formally correct without any assumption about $\underline{x}$, they are *of practical value only when* $\underline{x}$ is discrete.

The observed treatment assignments and covariates are z_i, $\underline{x}_i$, $i = 1,\ldots,n$. For conditions A, B, C, ... , let #(A, B, C, ...) be the number of vectors $(z_i, \underline{x}_i)$ which satisfy all of A and B and C and ... . For example, #(z=1, $\underline{x}$ = (1,0)) is the number of vectors $(z_i, \underline{x}_i)$ such that $z_i = 1$, $\underline{x}_i = (1,0)$. Define the sample conditional proportion $p(A|B)$ by

$$\text{prop}(A|B) = \frac{\#(A,B)}{\#(B)} \quad \text{if} \quad \#(B) \neq 0$$

and leave $\text{prop}(A|B)$ undefined if $\#(B) = 0$.

Estimate $e(\underline{x})$ by $\hat{e}(\underline{a}) = \text{prop}(z=1 \mid \underline{x} = \underline{a})$. If $\hat{e}(\underline{a}) = 0$ or 1 then all units with $\underline{x} = \underline{a}$ received the same treatment. Theorem 1', which parallels theorem 1, shows that at all intermediate values of $\hat{e}(\underline{a})$, that is for $\hat{e}(\underline{a}) \in (0,1)$, there is balance.

Theorem 1'. Suppose $\hat{e}(\underline{a}) \in (0,1)$. Then

$$\text{prop}(z=b, \underline{x} = \underline{a} \mid \hat{e}(\underline{x}) = \hat{e}(\underline{a}))$$
$$= \text{prop}(z=b \mid \hat{e}(\underline{x}) = \hat{e}(\underline{a}))\ \text{prop}(\underline{x}=\underline{a} \mid \hat{e}(\underline{x}) = \hat{e}(\underline{a})) \quad .$$

Proof: Since if $\#(B,C) \neq 0$, then

$$|\text{prop}(A|B,C) - \text{prop}(A|B)| \cdot \text{prop}(B|C)$$

$$= \left|\frac{\#(A,B,C)}{\#(B,C)} - \frac{\#(A,B)}{\#(B)}\right| \cdot \frac{\#(B,C)}{\#(B)}$$

$$= \left|\frac{\#(A,B,C)}{\#(B)} - \frac{\#(A,B)\#(B,C)}{\#(B)\#(B)}\right|$$

$$= |\text{prop}(A,C|B) - \text{prop}(A|B)\text{prop}(C|B)| \quad ,$$

it is sufficient to show

$$\text{prop}(z=1|\underline{x}=\underline{a}) = \text{prop}(z=1|\hat{e}(\underline{x}) = \hat{e}(\underline{a})) \quad .$$

Now,

$$\text{prop}(z=1|\hat{e}(\underline{x}) = c) = \sum_{\underline{d}:\hat{e}(\underline{d})=c} \text{prop}(z=1|\underline{x}=\underline{d})\text{prop}(\underline{x}=\underline{d}|\hat{e}(\underline{x}) = \hat{e}(\underline{d}))$$

$$= \text{prop}(z=1|\underline{x}=\underline{d}) \sum_{\underline{d}:\hat{e}(\underline{d})=c} \text{prop}(\underline{x}=\underline{d}|\hat{e}(\underline{x}) = \hat{e}(\underline{d}))$$

$$= \text{prop}(z=1|\underline{x}=\underline{d})$$

as desired. //

Similar theorems and proofs about sample balance parallel theorems 2 and 5. In particular, in parallel with theorem 5, if $\hat{e}^*(\underline{x})$ is close to $\hat{e}(\underline{x})$ for all $\underline{x}$, then there is nearly sample balance at each value of $\hat{e}^*(\underline{x})$. For example $\hat{e}^*(\underline{x})$ might result from a logit model which closely fits the sample data.

4. THREE APPLICATIONS OF PROPENSITY SCORES TO OBSERVATIONAL GROUPS

The general results that we have presented suggest that, in practice, adjustment for the propensity score should be an important component of analysis of observational studies, because evidence of residual bias in the propensity score is evidence of potential bias in estimated treatment effects. We conclude with three examples of how propensity scores can be explicitly used to adjust for confounding variables in observational studies. The examples involve the three standard techniques for adjustment in observational studies noted by Cochran (e.g., 1965) and summarized by Rubin (1981), namely, matched sampling, subclassification, and covariance adjustment.

4.1 The Use of Propensity Scores to Construct Matched Samples from Treatment Groups

Matching is a method of sampling from a large reservoir of potential controls to produce a control group of modest size in which the distribution of covariates is similar to the distribution in the treated group. Some sampling of the control reservoir is often required to control costs associated with measuring the response, for example, costs associated with extensive follow-up of patients in clinical studies.

Although there exist model-based alternatives to matched sampling (e.g. covariance adjustment on random samples), there are several reasons why matching is appealing.

1. Matched treatment and control pairs allow relatively unsophisticated researchers to immediately appreciate the equivalence of treatment and control groups, as well as to perform simple matched pair analyses which adjust for

confounding variables. This issue is discussed in greater detail below in subsection 4.2 on balanced subclassification.

2. Even if the model underlying a statistical adjustment is correct, the variance of the estimate of the average treatment effect (1.1) will be less in matched samples than in random samples since the distribution of $\underline{x}$ in treated and control groups is more similar in matched than in random samples. To verify this, inspect the formula for the variance of the covariance adjusted estimate (e.g. Snedecor and Cochran, 1978, p. 368), and note that the variance decreases as the difference between treatment and control means on $\underline{x}$ decreases.

3. Model based adjustment on matched samples is usually to be more robust to departures from the assumed form of the underlying model than model-based adjustment on random samples (cf. Rubin, 1973b, 1979), primarily because of the more limited reliance on the model and its extrapolation.

4. In studies with limited resources but large control resevoirs and many confounding variables, the confounding variables can often be controlled by multivariate matching, but the small sample sizes in the final groups do not allow control of all variables by model-based methods.

A multivariate matching method is said (Rubin, 1976a,b) to be equal percent bias reducing (EPBR) if the bias in each coordinate of $\underline{x}$ is reduced by the same percentage. Matching methods which are not EPBR have the potentially undesirable property that they increase the bias for some linear functions of $\underline{x}$. If matched sampling is performed before the reponse (r_1, r_2) can be measured, and if all that is suspected about the relation between (r_1, r_2) and $\underline{x}$ is that it is approximately linear, then EPBR matching methods are reasonable in that they lead to differences in mean response in matched samples that should be less biased than in random samples.

In section 2 we observed that discriminant matching is equivalent to matching on the propensity score if the covariates $\underline{x}$ have a multivariate normal distribution. Assuming multivariate normality, Rubin (1976a) showed that matching on the population or sample discriminant is EPBR. We now show that matching on the population propensity score is EPBR under weaker distributional assumptions. It is assumed that the matching algorithm matches each treated $(z = 1)$ unit with a control $(z = 0)$ unit drawn from a resevoir of control units on the basis of the propensity score.

For convenience write e for the propensity score $e(\underline{x})$. The initial bias in $\underline{x}$ is

$$\underline{b} = E(\underline{x} \mid z = 1) - E(\underline{x} \mid z = 0) \quad .$$

Suppose we have a random sample of treated (z=1) units and a large reservoir of randomly sampled control units, and suppose each treated unit is matched with a control unit from the reservoir. Then the expected bias in matched samples is

$$\underline{b}_m = E(\underline{x}|z=1) - E_m(\underline{x}|z=0) \quad ,$$

where the subscript m indicates the distribution in matched samples. Thus the reduction in bias of $\underline{x}$ due to matching is

$$\underline{b} - \underline{b}_m = E_m(\underline{x}|z=0) - E(\underline{x}|z=0) \quad . \tag{4.1}$$

Theorem 6.

For any matching method that uses e alone to match each treated unit (z=1) with a control unit (z=2), the reduction in bias is

$$\underline{b} - \underline{b}_m = \int E(\underline{x}|e)[p_m(e|z=0) - p(e|z=0)]de \quad . \tag{4.2}$$

Proof: From (4.1) we have

$$\underline{b} - \underline{b}_m = \int [E_m(\underline{x}|z=0,e)p_m(e|z=0) - E(\underline{x}|z=0,e)p(e|z=0)]de \quad . \tag{4.3}$$

For any matching method satisfying the condition of the theorem,

$$E_m(\underline{x}|z=0,e) = E(\underline{x}|z=0,e) \tag{4.4}$$

because any matching method using e alone to match units alters the marginal distribution of e in the control group ($z=2$), but does not alter the conditional distribution of $\underline{x}$ given e in the control group.

However, by Theorem 1,

(4.5) $$E(\underline{x}|z=0,e) = E(\underline{x}|e) \quad .$$

Substitution of (4.4) and (4.5) into equation (4.3) yields the result (4.2). //

Corollary: If $E(\underline{x}|e) = \underline{\alpha} + \underline{\beta}e$ for some vectors $\underline{\alpha}$ and $\underline{\beta}$, then matching on the propensity score e alone is EPBR.

Proof: The percent reduction in bias for the ith coordinate of $\underline{x}$ is, from equation (4.2),

$$\frac{\beta_i [E_m(e \mid z = 0) - E(e \mid z = 0)]}{\beta_i [E(e \mid z = 1) - E(e \mid z = 0)]}$$

which is independent of i, as required. //

Rubin's (1979) simulation study examines the small sample properties of discriminant matching in the case of normal covariates with possibly different covariances in the treatment groups, so the study includes situations where the true propensity score is a quadratic function of $\underline{x}$, but the discriminant score is a linear function of $\underline{x}$. Table 1 presents previously unpublished results from Rubin's (1979) study for situations in which the propensity score is a monotone function of the linear discriminant, so propensity matching and discriminant matching are effectively the same. The covariates $\underline{x}$ are bivariate normal with common covariance matrix I and bias B along the standardized population discriminant. In the simulation, fifty treated units are matched using nearest available matching (Cochran and Rubin (1973)) on the sample discriminant with 50 control units drawn from a resevoir of 50R potential control units, for $R = 2,3,4$; details are found in Rubin (1979).

Assuming parallel linear response surfaces, table 1 shows that even in the absence of additional adjustments, propensity (discriminant) matching alone can remove most of the bias if the resevoir is relatively large. Moreover, table 1 shows that the population and sample propensity scores are about equally effective in removing bias, so no substantial loss is incurred by having to estimate the propensity score. It should be noted that the conditions underlying table 1 differ from the conditions underlying theorem 1 in as much as nearest available matching provides only a partial adjustment for the propensity score since exact matches are not generally obtained.

Propensity matching should prove especially effective relative to Mahalanobis metric matching (Cochran and Rubin (1973), Rubin (1976a,b, 1979, 1980)) in situations where markedly nonspherically distributed $\underline{x}$ make the use of a quadratic metric unnatural as a measure of distance between treated and control units. For example, we have found in practice that if $\underline{x}$ contains one coordinate representing a rare binary event, then Mahalanobis metric matching may try too hard to exactly match that coordinate, thereby reducing the quality of matches on the other coordinates of $\underline{x}$. Propensity matching can effectively balance rare binary variables for which it is not possible to adequately match treated and control units on an individual basis.

Table 1.

Percent Reduction in Bias Due to Matched Sampling
Based on the Sample and Population Propensity Scores*

Ratio of Size of Control Resevoir to Size of Treatment Group	Propensity Score Used for Matching	Initial Bias (B) .25	.50	.75	1.00
2	Sample	92	85	77	67
	Population	92	87	78	69
3	Sample	101	96	91	83
	Population	96	95	91	84
4	Sample	97	98	95	90
	Population	98	97	94	89

* Assuming bivariate normal covariates with common covariance matrix, parallel linear response surfaces, sample size of 50 in treated and control groups. Estimated percent reduction in bias from Rubin's (1979) study. The largest estimated standard error for this table is less than .03.

4.2 Subclassification on Propensity Scores

A second major method of adjustment for confounding variables is subclassification, in which experimental and control units are divided on the basis of $\underline{x}$ into subclasses or strata (Cochran (1965, 1968), Cochran and Rubin (1973)). Direct adjustment with subclass total weights can be applied to the subclass differences in response to estimate the average treatment effect (1.1) whenever treatment assignment is strongly ignorable (theorem 4) without modelling assumptions such as parallel linear response surfaces.

As a method of multivariate adjustment, subclassification has the advantage that it involves direct comparisons of ostensibly comparable groups of patients within each subclasses and therefore can be both understandable and pursuasive to an audience with limited statistical training. The comparability of patients within strata can be verified by the simplest methods, such as bar charts of means. Since the results of observational studies are often disputed, and since such disputes are not always confined to statistically sophisticated participants and audiences, correct results should be presented in a manner which is both persuasive and understandable to the study's audience. Cox (1981, p. 291) emphasizes the importance of presenting results in ways that are "vivid, simple, and accurate." Of course, it should be stressed that balance on observed covariates $\underline{x}$ does not imply balance on unobserved covariates.

A major problem with subclassification, noted by Cochran (1965), is that as the number of confounding variables increases, the number of subclasses grows dramatically, so that even with only two categories per variable, yielding 2^p subclasses for p variables, most subclasses will not have both treatment and control units. Subclassification on the propensity score is a natural way to obviate this problem.

We now use an estimate of the propensity score to subclassify patients in an actual observational study of therapies for coronary artery disease. The treatments are coronary artery bypass surgery ($z = 1$) and drug therapy ($z = 0$). The covariates $\underline{x}$ are clinical, hemodynamic, and demographic measurements on each patient made prir to treatment assignment. Even though the covariates have quite different distributions in the two treatment groups, within each of the five subclasses, the surgical and drug patients will be seen to have similar sample distributions of $\underline{x}$.

The propensity score was estimated using a logit model for z given $\underline{x}$. Covariates and interactions among covariates were selected for the model using a stepwise procedure. Based on Cochran's (1968) observation that subclassification with five subclasses is sufficient to remove at least 90% of the bias for many continuous distributions, five subclasses of equal size were constructed at the quintiles of sample distribution of the propensity score, each containing 303 patients. Beginning with the subclass with the highest propensity scores, the five subclasses contained 234 surgical patients, 164 surgical patients, 98 surgical patients, 68 surgical patients and 26 surgical patients, repectively.

For each of the 74 covariates, table 2 summarizes the balance before and after subclassification. The column labeled "2-Sample" in table 2 contains F-statistics, that is the square of the usual two-sample t-statistics for comparing the surgical group and drug group means of each variable prior to subclassification. The last two columns of table 2 contain F-statistics for the main effect of treatment and for the interaction in a 2×5, treatments by subclasses analysis of variance, performed for each covariate. It is

Table 2a

F-TESTS OF BALANCE BEFORE AND AFTER STRATIFICATION

Variable	2-Sample	2-Way Anova	
		Main Effect	Interaction
1	4.4	0.0	0.7
2	18.1	0.0	0.7
3	6.8	0.0	1.4
4	25.0	0.2	0.8
5	5.3	1.0	0.9
6	7.3	2.2	1.2
7	26.0	0.2	0.3
8	10.9	1.6	0.5
9	11.6	1.2	1.0
10	6.8	0.1	1.2
11	38.4	0.4	1.4
12	9.0	0.1	2.9*
13	6.8	0.0	0.8
14	7.3	0.1	0.4
15	4.4	0.0	0.2
16	23.0	0.0	0.6
17	10.2	0.3	1.1
18	31.4	0.1	2.2
19	4.8	0.1	0.7
20	6.2	0.1	0.9
21	20.2	0.2	1.3
22	7.8	0.5	0.9
23	10.2	0.6	0.8
24	4.8	0.2	0.0
25	6.8	0.0	1.3
26	25.0	0.2	0.0
27	10.9	0.2	0.3
28	10.9	0.2	0.2
29	4.0	0.0	1.3
30	5.8	0.1	0.1
31	8.4	0.3	0.5
32	13.0	0.1	0.2
33	13.0	2.1	0.4
34	16.0	0.1	1.4
35	24.0	0.3	0.1
36	16.0	1.0	0.2
37	9.6	0.7	0.4
38	10.9	0.7	0.2
39	4.0	0.2	0.8
40	14.4	0.1	0.4
41	7.8	0.7	0.8
42	51.8	0.4	0.9
43	14.4	0.1	0.4
44	9.6	1.0	1.3
45	29.2	0.3	0.4
46	4.3	0.5	0.8
47	18.5	0.3	2.2
48	7.8	0.4	0.5
49	15.2	0.4	0.2

Table 2b

F-TESTS OF BALANCE BEFORE AND AFTER STRATIFICATION

Variable	2-Sample	2-Way Anova	
		Main Effect	Interaction
50	5.8	0.0	0.8
51	19.4	0.0	0.3
52	5.8	2.3	1.4
53	13.0	3.6	2.0
54	6.2	0.6	0.8
55	8.4	0.9	0.9
56	16.0	1.1	0.4
57	5.8	1.6	0.8
58	6.8	0.2	2.4
59	4.8	0.5	0.9
60	14.4	0.2	0.6
61	7.8	0.0	1.1
62	22.1	0.3	0.2
63	6.2	1.0	1.2
64	11.6	0.2	0.3
65	18.5	0.3	0.2
66	43.6	0.1	1.4
67	31.4	0.0	1.0
68	18.5	0.0	0.7
69	13.0	0.8	2.3
70	10.9	0.0	2.1
71	10.9	0.0	2.4
72	11.6	0.4	1.4
73	16.8	0.0	1.2
74	7.8	3.1	0.5

easily seen that there is considerable imbalance prior to subclassification, and yet within subclasses there is greater balance than would have been expected if treatments had been assigned at random within each subclass.

Subclassification on the propensity score is not the same as any of the several methods proposed by Miettinen (1976): the propensity score is not generally a "confounder" score. (For example, one of Miettinen's confounder scores is $p(z=1|r_z=1,\underline{x}) \neq p(z=1|\underline{x}) = e(\underline{x})$.)

4.3 Propensity Scores and Covariance Adjustment

The third standard method of adjustment in observational studies is covariance adjustment. The point estimate of the treatment effect obtained from analysis of covariance adjustment for multivariate $\underline{x}$ is, in fact, equal to the estimate obtained from univariate covariance adjustment for the sample linear discriminant based on $\underline{x}$, whenever the same sample covariance matrix is used for both the covariance adjustment and the discriminant analysis. This fact is most easily demonstrated by linearly transforming $\underline{x}$ to $(d,\underline{v})$ where d is the sample discriminant, and $\underline{v}$ is orthogonal to the sample discriminant and thus has the same sample mean in both groups. Since covariance adjustment is effectively adjustment for the linear discriminant, plots of the responses r_{1i} and r_{0i} or residuals $r_{ki} - \hat{r}_{ki}$ (where $\hat{r}_{ki}$ is the value of r_{ki} predicted from the regression model used in the covariance adjustment) vs the linear discriminant are useful in identifying nonlinear or nonparallel response surfaces, as well as extrapolations, which might distort the estimate of the average treatment effect. Furthermore, such a plot is a bivariate display of multivariate adjustment, and as such might be useful for general presentation.

Generally, plots of responses and residuals from covariance analysis against the propensity score $e(\underline{x})$ are more appropriate then against the discriminant, unless of course the covariates are multivariate normal with common covariance matrix in which case the propensity score is a monotone function of the discriminant. The reason is that if treatment assignment is strongly ignorable, at each $e(\underline{x})$ the expected difference in response $E(r_1 \mid z=1, e(\underline{x})) - E(r_0 \mid z = 2, e(\underline{x}))$ equals the average treatment effect at $e(\underline{x})$, namely $E(r_1|e(\underline{x})) - E(r_0|e(\underline{x}))$. This property holds for the propensity score $e(\underline{x})$ and for any balancing score $b(\underline{x})$, but does not generally hold for other functions of $\underline{x}$; generally, plots against other functions of $\underline{x}$ are still confounded by $\underline{x}$. Consequently, a plot of the responses r_1, r_2 or residuals against $e(\underline{x})$ can reveal particularly important nonparallelism, nonlinearity, or extrapolations in the response surfaces. Since the purpose of such a plot is to reveal departures from assumptions, some enhancement to accentuate trends in the plot will often be necessary using, perhaps, techniques such as described by Cleveland and Kleiner (1975).

Cases where covariance adjustment has been seen to perform quite poorly are precisely those cases in which the linear discriminant is not a monotone function of the propensity score, so that covariance adjustment is implicitly adjusting for a poor approximation to the propensity score. In the case of univariate x, the linear discriminant is a linear function of x, whereas the propenisty score may not be a monotone function of x if the variances of x in the treated and control groups are unequal. (Intuitively, if the variance of x in the control group is much larger than the variance in the treated group, then individuals with the largest and smallest x values usually come from the control group.) Rubin (1973b, tables 4 and 6, with

r=1, and $\hat{\tau}_p$ as the estimator) has shown that univariate covariance adjustment will either increase the bias by up to 304% or overcorrect for bias by 298% for certain exponential response surfaces if the variances of x in the treated and control groups differ by a 2:1 ratio. Unequal variances of covariates are not uncommon in observational studies, since the subset of units which receives a new treatment is often more homogeneous than the general population. For example, in the observational half of the Salk Vaccine trial, the parents of second graders who volunteered for vaccination had higher and therefore less variable educational achievement (x) than parents of control children, that is parents of all first and third graders (Meier (1972)).

In the case of multivariate normal $\underline{x}$, Rubin (1979, table 2) has shown that covariance adjustment can increase the expected squared bias by as much as 55% if the covariance matrices in the treated and control groups are unequal; that is, if the discriminant is not a monotone function of the propensity score. In contrast, when the covariance matrices are equal, so the discriminant is a monotone function of the propensity score, covariance adjustment removes between 84% and 100% of the expected squared bias in the cases considered by Rubin (1979, table 2).

REFERENCES

Cleveland, W. S. and Kleiner, B. (1975). A graphical technique for enhancing scatterplots with moving statistics. Technometrics, 17:447-455.

Cochran, W. G. (1965). The planning of observational studies of human populations. Journal of the Royal Statistical Society, Series A, 234-255.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics, 205-213.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: a review. Sankhya, Series A, 417-446.

Cox, D. R. (1958). The Planning of Experiments. New York: John Wiley and Sons.

Cox, D. R. (1970). The Analysis of Binary Data. London: Methuen.

Cox, D. R. (1972). The analysis of multivariate binary data. Applied Statistics, 21:113-120.

Cox, D. R. (1981). Theory and general principle in statistics. Journal of the Royal Statistical Society, Series A, 144:289-297.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). Journal of the Royal Statistical Society, Series B, 41:1-31.

Dempster, A. P. (1971). An overview of multivariate data analysis. Journal of Multivariate Analysis, 1:316-346.

Fisher, R. A. (1953). The Design of Experiments. London: Hafner.

Kempthorne, O. (1952). The Design and Analysis of Experiments. New York: John Wiley and Sons.

Meier, P. (1972). The biggest public health experiment ever: The 1954 trial of the Salk poliomyelitis vaccine. In J. M. Tanur, et. al., (eds.) Statistics: A Guide to the Unknown. San Francisco: Holden Day.

Miettinen, O. (1976). Stratification by a multivariate confounder score. American Journal of Epidemiology, 104:609-620.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. Biometrics, 29:159-183. Printer's correction note 30:728.

Rubin, D. B. (1973b). The use of matching and regression adjustment to remove bias in observational studies. Biometrics, 29:184-203.

Rubin, D. B. (1976). Matching methods that are equal percent bias reducing: Some examples. Biometrics.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics, 2:1-26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. Annals of Statistics, 6:34-58.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association, 74:318-328.

Rubin, D. B. (1980). Bias reduction using Mahalanobis metric matching. Biometrics, 36:293-298.

Rubin, D. B. (1981). William G. Cochran's contributions to the design and analysis of observational studies. To appear in Memorial Book edited by Rao and Sedransk.

Snedecor, G. W. and Cochran, W. G. (1980). Statistical Methods. Ames, Iowa: Iowa State University Press.

PRR/DBR/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2305	2. GOVT ACCESSION NO. AD-A114 514	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) The Central Role of the Propensity Score in Observational Studies for Causal Effects		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Paul R. Rosenbaum and Donald B. Rubin		8. CONTRACT OR GRANT NUMBER(s) HRA 230-76-0300 DAAG29-80-C-0041 P30-CA-14520
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS (see Item 18 below)		12. REPORT DATE December 1981 13. NUMBER OF PAGES 28
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Cancer Institute
9000 Rockville Pike
Bethesda, MD 20205

Health Resources Administration
3700 East-West Highway
Hyattsville, MD 20782

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Observational studies; covariance adjustment; subclassification; matching

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The results of observational studies are often disputed because of nonrandom treatment assignment. For example, patients at greater risk may be overrepresented in some treatment group. This paper discusses the central role of "propensity scores" and "balancing scores" in the analysis of observational studies. The propensity score is the (estimated) conditional probability of assignment to a particular treatment given a vector of observed covariates. Both large and small sample theory show that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Applications ...

(continued)

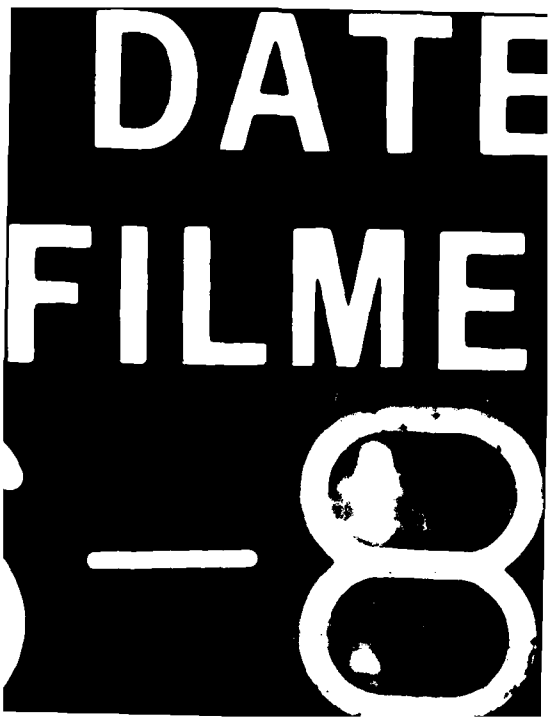
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT (continued)

include: (1) matched sampling on the univariate propensity score which is equal percent bias reducing under more general conditions than required for discriminant matching, (2) multivariate adjustment by subclassification on balancing scores where the same subclasses are used to estimate treatment effects for all outcome variables and in all subpopulations, and (3) visual representation of multivariate adjustment by a two-dimensional plot.


DATE
FILME
8